BIO STATISTICS COLLABORATION OF AUSTRALIA

Study Guide

Data Management and Statistical Computing (DMC)

Semester 1, 2022

Prepared by:

Murthy Mittinty
School of Public Health
The University of Adelaide

David Fitzgerald
School of Public Health
The University of Queensland

Copyright © The University of Queensland, The University of Adelaide

THE UNIVERSITY OF ADELAIDE AUSTRALIA

THE UNIVERSITY OF QUEENSLAND

Contents

Data Management and Statistical Computing (DMC)

Semester 1, 2022

Contact details

Murthy N Mittinty

School of Public Health
The University of Adelaide
AHMS Building
57 North Terrace, Adelaide 5000
(08) 8313 0961
murthy.mittinty@adelaide.edu.au

If you have any general BCA queries, please contact: Karolina Kulczynska-Le Breton or Emily Higginson at the BCA Coordinating Office on 02 9562 5076/54 or via email to bca@sydney.edu.au

Background

The aim of this unit is to provide students with the knowledge and skills required to undertake moderate to high-level data manipulation and management in preparation for statistical analysis of data typically arising in health and medical research.

Unit summary

Data comes in various shapes and sizes. The data that we use for analysis can come from various sources, these include public use data sets, administrative data, registries, and hospital data, surveys and census data. Some of these data maybe well managed within a data custodian agency. These data can be stored in various formats like excel sheet, can be in stored in Stata, or any other data formats. Additionally, these might have good data dictionaries which can be used readily. However, not every data set maybe clean. Such unclean messy data needs to be cleaned. Data dictionaries to be developed. Data dictionaries refers to the part of defining the values of a variable, range of values of a variable, defining labels for every value that is in the data set, especially for categorical variables. Defining consistent names for each variable across different data sets, when using multiple data sets from various sources. Once the data are clean, they are ready for analysis. The first steps of data analysis include data visualisation, data description and complex data analysis. This course will teach you the basics like importing and exporting data, recording and formatting data, developing proper dictionaries. Once the data are exported or imported into a software that the analyst will be comfortable analysing. Then, the next step is data visualisation, this includes creating histogram, bar plots, pie charts, and more complex data visuals. This course will show you how this aspect can be achieved when using Stata or R. The logical next step following data visualisation is data analysis. This includes taking summaries, writing user defined functions, preparing publication quality tables, generating new variables from existing variables (e.g., creating BMI for weight and height). All these and many more interesting topics will be covered in the three modules over the coming 13 weeks.

Workload requirements

The expected workload for this unit is 10-12 hours per week on average, consisting of guided readings, discussion posts, independent study and completion of assessment tasks.

Prerequisites

None

Co-requisites

None

Learning Outcomes

At the completion of this unit students should be able to:

1. undertake data manipulation and management using two major statistical software packages (Stata and R);
2. appropriately display and summarise data using statistical software;
3. understand how to check and clean data;
4. link data files through unique and non-unique identifiers;
5. have fundamental programming skills for efficient use of statistical software;
6. understand key principles of confidentiality and privacy in data storage, management and analysis.

Unit content

The unit is divided into 3 modules, summarised in more detail below. Each module will involve approximately 4 weeks of study and generally includes the following material:

1. Module notes describing concepts and methods, and including some exercises of a more "theoretical" nature.
2. Selected readings from published articles or textbooks.
3. One or more extended examples illustrating the concepts/methods introduced in the notes and including more practically oriented exercises.

Study materials for all Modules are downloadable from the eLearning unit site. Assignments and supplementary material, such as datasets will be posted to the unit site. Please note that we are not able to post copies of copyright material (journal articles and book extracts)—for these you will have to rely on the hard copy mail-out or resources from your home university's library.

Recommended approaches to study

Students should work through each module systematically, following the module notes and any readings referred to, and working through the accompanying exercises. *You will learn a lot more efficiently if you tackle the exercises systematically as you work through the notes.* You are encouraged to post any content-related questions to eLearning, whether they relate directly to a given exercise, or are a request for clarification or further explanation of an area in the notes. You should also work through all of the computational examples in the notes for yourself on your own computer.

Outline solutions to the exercises in each module (except those to be submitted for assessment, as described below) will be posted.

Method of communication with coordinator(s)

Questions about administrative aspects or course content can be emailed to the coordinator, and when doing so please use “DMC:” in the Subject line of your email to assist in keeping track of our email messages. Coordinator/s will be available to answer questions related to the module notes and practical exercises, and to address any other issues that require clarification. However, please note that instructors are not necessarily available every day of the week and you should expect that it may take a day or so to respond to questions (possibly longer over weekends and during breaks!).

We strongly recommend that you post content-related questions to the Discussions tool in the DMC area of BCA’s eLearning site. In 2021 we are using the Learning Management system hosted by the University of Sydney. You may be familiar with the system from previous BCA units, and will receive any specific instructions on using the eLearning site this semester from the BCA Coordinating Office. There is also a “Getting Started” document available on the Student Resources page of the BCA website.

Module descriptions

There are 3 modules in this course; each module has been divided into Part A and Part B, each scheduled over a fortnight. Each module sub-section is scheduled to begin on a Monday and conclude on the Sunday of the following week. The due date for submission of required assignments from each module is 2:00 pm (Australian Eastern Standard Time) on the due date.

Below is an outline of the study modules, followed by a timetable and assessment description table.

- Module 1: The basics. Importing and exporting data; recoding and formatting data; labelling variables and values; use of date data, displaying and summarising data.

- Module 2: Graphs, Data management and Statistical Quality Assurance Methods. Includes advanced graphics for production of publication-quality graphs.

- Module 3: More Advanced Statistical Computing: Using functions to generate new variables; appending, merging and transposing data; programming skills including loops, arguments and programs/macros.

Unit schedule

Semester 1, 2022 starts on Monday 28 February

Week	Week commencing	Module	Topic	Assessment
1	28 February	Module 1A	Basic introduction to Stata	
2	7 March	Module 1 A	Basic introduction to Stata	Assignment 1 available 11th March
3	14 March	Module 1B	Basic introduction to R	
4	21 March	Module 1B	Basic introduction to R	
5	28 March	Module 2A	Graphics Basics in Stata and R	**Assignment 1 due 4th April**
6	4 April	Module 2A	Graphics Basics in Stata and R	
7	11 April	Module 2B	Data Management and Exploratory Data Analysis	Assignment 2 Available on 15th April
Mid semester break 18 April				
8	25 April	Module 2B	Data Management and Exploratory Data Analysis	
9	2 May	Module 2c	Supplementary Data Visualisation	Assignment 2 due 6th May
10	9 May	Module 3A	Advanced Data Management and Statistical Computing Using Stata	
11	16 May	Module 3A	Advanced Data Management and Statistical Computing Using Stata	Assignment 3 Available 16th May
12	23 May	Module 3B	Advanced Data Management and Statistical Computing Using R	
13	30 May	Module 3B	Advanced Data Management and Statistical Computing Using R	
	6 June		Work on Assignment 3	**Assignment 3 due 9th June**

Assessment

Assessment will include three written assignments worth 30%, 35% and 35% each, to be made available 2.5 weeks before the due date. These assignments will be posted on the eLearning site together with an online Announcement broadcasting their availability. Individual feedback will be provided to each student; model solutions will also be provided once all marked assignments have been returned. Summary statistics on results for the entire class will also be provided. Assignments should be submitted via the assignment submission tool on eLearning; if you experience difficulties with this submission method, assignments can be submitted via email.

Assessment name	Assessment type	Coverage	Learning objectives	Weight
Major Assignment 1	Assignment	Modules 1A-B	1,2	30%
Major Assignment 2	Assignment	Modules 2A-B	1,2,3,4	35%
Major Assignment 3	Assignment	Modules 1A-3B	1,2,3,4,5,6	35%

In general, you are required to submit work typed in Word or similar. We strongly recommend you become familiar with equation typesetting software such as Microsoft's Equation Editor for algebraic work. You may submit neatly handwritten work, however please note that marks will potentially be lost if the solution cannot be understood by the markers due to unclear or illegible writing. Handwritten work should be scanned and collated into a single pdf file and submitted via the eLearning site. See the BCA Assessment Guide for guidelines on acceptable standards for assessable work.

Students are encouraged to discuss relevant topics in the Discussion Board. However, please avoid posting questions relating directly to assessable material. These should be emailed to the Unit Coordinator in the first instance.

Explicit solutions to assessable exercises should not be posted for others to use. Each student's submitted work must be clearly their own, with anything derived from other students' discussion contributions clearly attributed to the source.

Submission of assessments and academic honesty policy

All assessment material should be submitted via the relevant Assessment module in Canvas unless otherwise advised. Turnitin plagiarism detection is applied to all submissions. For detailed information, please see the BCA Assessment Guide, which includes links to the Academic Honesty policies at member universities. Please familiarise yourself with the procedures and policies at your home university. You will need to indicate your compliance with the plagiarism guidelines and policy at your home university.

A special note regarding "contract cheating" sites: Unfortunately there have been instances in the past of students using such websites to post assignment questions and receive solutions (usually for a fee). We have arrangements with these sites to identify the students posting questions or accessing the solutions, and such students will be referred to and face disciplinary processes at their home university.

Late submission of assessments and extension procedure

The standard BCA policy for late penalties for submitted work is a 5% deduction from the earned mark for each day the assessment is late, up to a maximum of 10 days (including weekends and public holidays). Extensions are possible, but these need to be applied for (by email) as early as possible. The Unit Coordinator can approve extensions up to three days; for extensions beyond three days, you must apply to your home university, using their standard procedures.

Learning resources

It is recommended that you have access to the following textbooks:
Jull S, Frydenberg M. An Introduction to Stata for Health Researchers, 4th ed. Stata Press, 2014.

Wickham H, Grolemund G. R for Data Science. O'Reilly 2017. Dalgaard, P. (available online https://r4ds.had.co.nz/).

Your University Library may have an ebook (Full Text Online) version of the Juul text; the Wickham text is freely available at the web link provided. If you have any issues accessing these texts please contact me.

Software requirements and assumed knowledge

For this subject you will need to have access to the following software packages:

- Stata version 12 or later (the latest version is v16)
- R version R64 3.4.2 or later (the latest version is 4.0.2)
- RStudio version 1.3 or later (the latest version is 1.4)

For help with R, please see Learning R in the Student Resources site.

If you have not yet organised access to these packages, you should do so as soon as possible. This is a practical course which requires regular use of the relevant software; delays in gaining access to these packages may impact your ability to complete the course. Information on how to download R and RStudio, and access Stata can be found in the BCA Textbook and Software Guide.

Feedback

Our feedback to you:

The types of feedback you can expect to receive in this unit are:

- Formal individual feedback on submitted exercises assignments
- Responses to questions posted on Canvas

Your feedback to us:

One of the formal ways students provide feedback on teaching and their learning experience is through the BCA student evaluation survey at the end of each semester. The feedback is anonymous and provides the BCA with evidence of aspects that students are satisfied with and areas for improvement.

Unit changes, including response to recent student evaluation

DMC was last delivered in Semester 2 2021. In Semester 1, 2020, the R notes were redeveloped by Dr Jennie Louise, and further changes, including incorporation of online tutorials and video content, was implemented in Semester 2, 2021. Further changes have been implemented in Semester 1, 2021, including additional modification of the course notes and incorporation of optional supplementary content.

Acknowledgments

Earlier versions of DMC course notes prepared by Jelena Romaniuk, Centre for Molecular, environmental, Genetic and Analytic Epidemiology, School of Population Health, The University of Melbourne; and Patrick McElduff & Catherine D’Este, Centre for Clinical Epidemiology and Biostatistics, Faculty of Health & Medicine, University of New Castle, Australia. Updated with Major revisions in 2017 by Jennie Louise, Faculty of Health Sciences, The University of Adelaide. Replacement of SAS with R in 2019 by Michael Waller, School of Public Health, The university of Queensland, Murthy N Mittinty, School of Public Health, The University of Adelaide and Jennie Louise, Faculty of Health Sciences, The University of Adelaide. Major revisions of Module 2 and 3, and R notes rewritten in 2020 by Jennie Louise.